Neuromorphic Event-based Space-Time Template Action Recognition

Neuromorphic Event-based Space-Time Template Action Recognition

S. Harrigan[1], D. Kerr[1], S.A. Coleman[1], P. Yogarajah[1], Z. Fang[2], C. Wu[2]

[1]*Intelligent Systems Research Centre, Ulster University, Northern Ireland, United Kingdom*
[2]*College of Information Science and Engineering, Northeastern University, Shenyang, China*
{harrigan-s; d.kerr, sa.coleman, p.yogarajah}@ulster.ac.uk. {fangzheng,wuchengdong}@mail.neu.edu.cn

Abstract

Neuromorphic vision hardware enables observed actions to be reduced to a series of spike trains; these trains contain unique properties relevant to observed actions. This paper presents an approach to event-based image processing which allows for the detection of specific fine grain actions through the adoption of template matching alongside neuromorphic hardware. The proposed approach was applied to the detection of breathing actions in an ambient assisted living (AAL) environment, this involved the detection of shallow, normal and heavy breathing for multiple participants using a single template. The results gained suggest that this approach could be useful when deployed in an AAL environment.

Keywords: event-based, neuromorphic image processing, space-time, ageing-in-place, action detection

1 Introduction

One of the most efficient designers of hardware is the evolutionary process, and the study of this process through biology has led to key insights into some of the most important aspects of biological engineering; these insights have been applied since the 1980s in the field now aptly labelled neuromorphic engineering. The field of neuromorphic engineering is a combination of biophysics, neuroscience, computer science and computer engineering. The primary goal of the field is the design, manufacturing and application of hardware and algorithms which mimic the neural system activity found in biological life.

A recent advancement in the field has been the development of vision sensor hardware which mimics the retina found in many animals; legacy vision sensors produce a series of frames (see Figure 1a and Figure 1c) which contain redundant information (constant capturing of frames regardless of any changes occurring in the scene), and an inherently high level of latency due to limitations in legacy sensor design. A neuromorphic vision sensor, such as iniLabs' Dynamic Vision Sensor (DVS), does not suffer from the same redundancy and latency issues as legacy vision sensors. The DVS mimics the human eye through the use of a silicon retina which logarithmically scales to the luminance observed at the individual pixel level (see Figure 1b and Figure 1d). If a change of luminance is detected by a sensor at a specific pixel location then an activation event is produced relating to that individual pixel, and is reported by DVS, if no change is detected then no activation event is produced which means that the DVS produces a sparse dataset containing data points corresponding only to locations at which an activity occurs in the observed scene. The human vision system has an approximate inherent latency of 180ms [Willey, 1985, Rayner et al., 2009], legacy vision sensors have approximately 33.4ms [Eberlein, 2015], while the DVS has been recorded with an inherent latency of approximately 0.015ms [Lichtsteiner, et al., 2008]. The advantage of this speed difference in neuromorphic vision sensors was illustrated by [Moeys, et al., 2016] where neuromorphic hardware was used for fast detection of a prey in a predator-prey scenario.

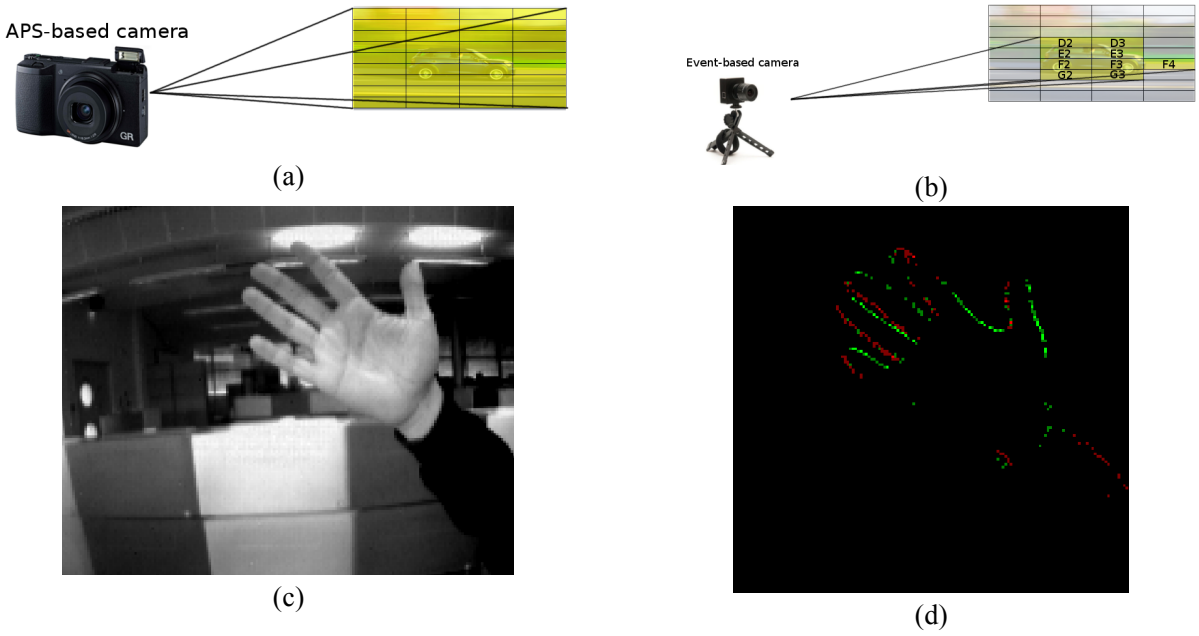

Figure 1: A representation of the image capture plane for both APS-based (a) and event-based (b) vision hardware, each cell in the plane represents a conceptual pixel; yellow cells indicate pixels where the luminance is captured. (c) An example frame captured by an APS-based hardware showing a hand waving, (d) the same observed motion of a waving hand observed by an event-based hardware, green represents an increase in light intensity while red indicates a decrease in light intensity, the labelled cells (D2, D3 etc.) are used to further illustrate that specific cells are activated instead of every cell.

The image processing field specialising in using neuromorphic vision sensors is referred to as event-based, or neuromorphic, image processing where the luminance change triggers an event for the respective pixel as discussed previously. Within this field, external legacy vision sensors are referred to as active-pixel sensors (APS) due to their unifying approach to luminance capture with all sensor values in the pixel array being captured during a single frame output. The ability to detect changes in a scene is an area of ongoing research in the field of APS-based image processing but within the subfield of event-based image processing, this is a trivial by-product of neuromorphic sensors. This is due to the definition of a scene change in both approaches which can be summarised as the change of luminance within the observed scene respective to the sensor. While APS-based approaches can suffer from noise (e.g. latency-induced capturing motion causing blurring), event-based vision sensors only report changes in luminance levels resulting in less noise in stable scenes and the datatype produced has the potential to enable new approaches to noise reduction to be explored.

2 APS-based Space-Time Templating approach

A frame-based approach to the detection of actions as they develop over time using a template was introduced by [Shechtman, et al., 2005]; this approach correlates an activity in space-time using a template. This approach utilises the induced motion field of an action under the assumption that the same action will result in the same induction within the motion field independent of the observed party. The calculation of the induced motion field acts as a descriptor for the observed action. The motion fields are representative of the intensity pattern induced by motion and can be correlated against other motion fields to determine similarity. This approach is illustrated in Figure 2 where a small space-time template T is correlated against a segment S of a larger video stream V along three-dimensions of x, y and t by sliding T along the three-dimensions.

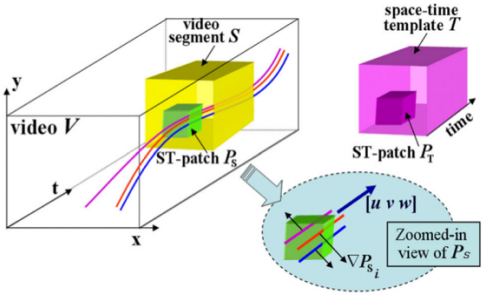

Figure 2: Frame-based space-time templating [Shechtman, et al., 2005, Figure 1]

This approach is capable of identifying multiple instances of an action across the video by comparing a patch P_S of S from the video against a patch P_T of T. This approach is invariant to appearance of objects and their background. Additionally, the approach requires no background-foreground segmentation. The approach is not invariant to geometric deformations such as scale and orientation except in small scale and orientation changes and it is very computationally expensive to perform. The proposed approach outlined in this paper, and its corresponding implementation, is inspired by this frame-based approach in terms of its sliding window system and template-to-video stream comparison methods, however, the proposed approach uses a less computationally expensive comparison method.

3 Proposed Neuromorphic-based Space-Time Templating approach

Recent experimentation in biological vision on the neuronal level has presented evidence that the retina reports observed scenes as a series of spikes (ON or OFF signals) [Thorpe, et al., 2001] known as a spike train (Figure 3); neuromorphic hardware, such as DVS, report events in a similar manner. An assumption of the proposed approach is that an action will result in the same spike train each time while a different action will result in a less similar spike train. The proposed approach is comprised of three stages: event-correlation, descriptor calculation and comparison-calculation. Each of these stages will be dealt with within this section.

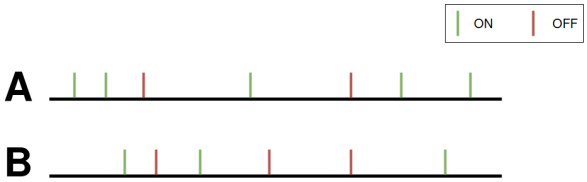

Figure 3: An illustration of two spike train, green spikes represent an ON signal and red spikes represent an OFF signal

Figure 3 shows an example illustration of two spikes train as they could be presented by neuromorphic hardware, with green representing an ON value and red representing an OFF value respectfully.

3.1 Event-Correlation

Determining the relevance of one event to another event is a large challenge in event-based image processing as the data obtained is spatial-temporal and is presented in a sequence (Figure 2) which can be a false representation of the order of activations of sensors from the arbitration process of the neuromorphic hardware. The approach outlined in this paper works under the assumption that events are spatial-temporal close within a small neighbourhood which would mitigate some of the correlation issues, but there is no falsehood that there is no margin of error in this strategy. For example, an event could be part of a different action occurring at the same time in the same region of the observed scene.

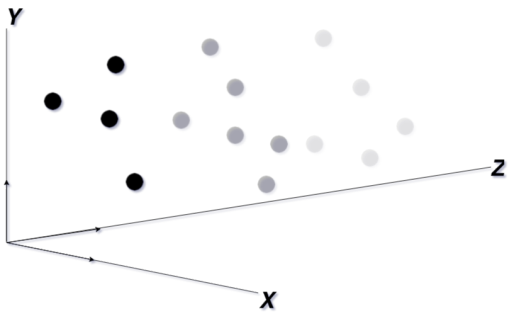

Figure 3: An illustration of an action in space-time for event-based hardware broken up into three groups which were detected at different times, black dots indicate most recent events, dark-grey temporally less prominent events and light grey even older events

In the proposed approach, a spatial-temporal map of all recent events (including *x* and *y* coordinates and time) is maintained. When a new event is processed this map is updated with the new events details.

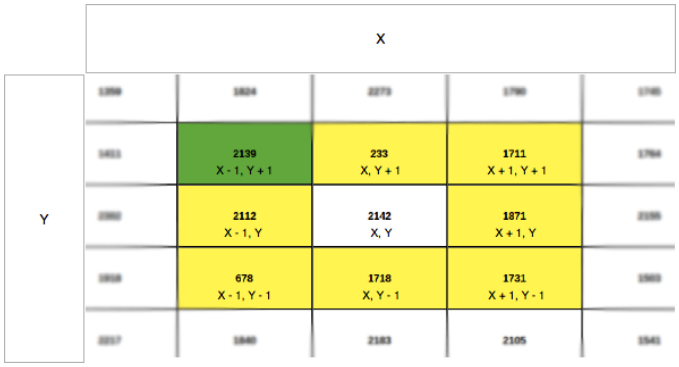

Figure 4: An Illustration of the temporal search performed by a 3x3 spatial-temporal correlator showing, the time in ms, of the last event occurring, centre white is the currently received event, green is the identified closest temporal activity.

Using a spatial-temporal 3x3 neighbourhood (Figure 4), centred on the current event being processed, to identify the closest event by time (within a time limit *TL*, which is typically set at 5ms); working under the assumption that relevant events to the current event will be within the same region and will be close in time. Once a neighbour is identified, this information is passed to the descriptor generators respectfully. Events which have no neighbours meeting the limit *TL* cause the neighbourhood to expand and the distance *D* (Equation 1) between events is calculated until distance limit *DL* (usually 10 pixels) is reached. If no corresponding event pixel is found then the event is added to the map, but nothing is passed to the descriptor calculations (section 3.2). The distance *D* is calculated as

$$D = \sqrt{k^2 + j^2} \tag{1}$$

where k is the difference between the correlated neighbour event and the current event's x coordinate, j is the difference between the correlated neighbour event and the current event's y coordinate.

3.2 Descriptor Calculation

The proposed approach generates two descriptors for both a Template T (an action being searched for) and a segment V_s of a video stream V, where V_s is acquired through a sliding window process. The two descriptors are angle-orientation and pattern respectfully. The angle-orientation descriptor is a series of bins which represent each angle ranging from 0^o to 359^o; each bin is incremented for the occurrence of its representative angle. This descriptor can be used to examine and compare the motions within an observed scene. As discussed in Section 3.1, the process of

event correlation provides the x, y and time values of the current event along with its correlated neighbour. The calculation of the angle is given by:

$$\theta = \tan^{-1}(\frac{v}{b}) \tag{2}$$

where b is the difference between the correlated neighbour event and the current event's x coordinate, v is the difference between the correlated neighbour event and the current event's y coordinate. The pattern descriptor describes the pattern of events observed, the pattern descriptor is a record of the events correlated b occurrence in time as discussed in Section 3.1.

3.3 Comparison-Calculation

The angle-orientation T_A and pattern descriptors T_P of T are calculated once and remain unchanged throughout the operation. Each new V_S of V possesses an angle-orientation V_A and pattern V_P descriptor as the sliding window proceeds along the stream. Equation 4 was developed to measure the correlation between T_A against V_A:

$$R_A = \lim_{i \to n} \left(\sum_{i=0}^{n} \frac{W_i}{T_i + \varepsilon} \right) = 1.0 \tag{4}$$

where n is the number of angles in the orientation descriptor, W_i is the value of the i-th bin occurring for V_S, T_i is the value of the corresponding i-th bin for T; ε is set appropriately to negate divisions by zero. Equation 5 was developed to measure the correlation between T_P against V_P:

$$R_P = \frac{\sum_{i=0}^{n} \begin{cases} 0 & M_i \neq Y_i \\ 1 & M_i = Y_i \end{cases}}{n} \tag{5}$$

where n is the number of events, Y_i is the value of the polarity of the i-th event in V_S, M_i is the value of the polarity of the i-th event in T. To calculate the overall similarity (a normalised metric which represents how closely correlated V_S is compared to T) between the results produced by Equation 4 and 5, Equation 6 was developed to produce a numeric representation of the similarity between T and V_S:

$$C = \frac{R_A \cdot R_P}{1 - (\min(R_A, R_P) + \varepsilon)} \tag{6}$$

where ε is set appropriately to negate divisions by zero, R_A and R_P come from Equation 4 and 5 respectfully.

4 Experimentation Results

The proposed approach was implemented and deployed in a controlled environment, with the goal being the detection of breathing via its underlying respiration motions. The implemented template T was a single recording of respiration motions of a control participant, but the intermediate pause (the period between inhalation and exhalation) was removed to minimise false readings. T had a temporal length of 44072 microseconds and contained over 61136 spatial-temporal events. It should be noted that at no point during the experiment was T changed. The benchmark for success in this experiment was the majority detection of breaths performed by a participant, a single breath was defined as both the inhalation and exhalation actions formed in sequence with a similarity score (Equation 6) of each motion rising above L which was set to 0.75 during the experiment. The DVS128 was the capturing hardware used throughout the experiment, Figure 5 shows an example of event-based hardware capturing some of the activity of an exhalation action with green pixels indicating an increase in the luminance detected by the neuromorphic hardware (the area observed has become brighter) and red pixels indicating and decrease in the luminance level (the area observed has become darker). ε (found in Equation 4 and Equation 6) was set to $1x10^{-7}$.

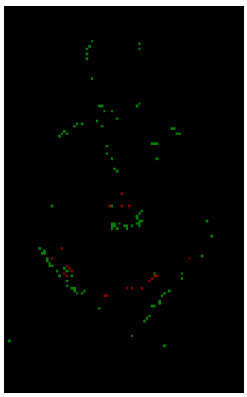

Figure 5: event-based hardware's representation of exhalation, using the same colour coding scheme as described in Figure 1 (d)

By calculating *C* (Equation 6), respiration motions can visually be observed. Figure 6 shows a sample plot of *C* over a 10 second period containing three breaths, the grouping of results illustrate both the inhalation, intermediate pause and exhalation.

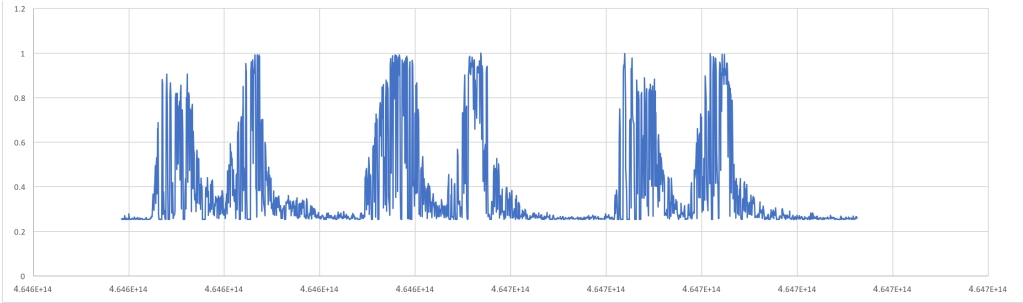

Figure 6: sample plot of *C* illustrating different phases of breathing with the x-axis representing time and the y-axis representing the value calculated for C

The experiment was separated into two phases, detection and distance. During the detection phase, four participants were positioned 0.5 metres from the capturing hardware while the implemented approach was monitoring their breathing actions. This was repeated for three separate runs, with normal, heavy and shallow breathing respectively while template *T* remained unchanged. The results of the distance phase are provided in Table 1, Table 2 and Table 3. Each table contains six entries, a control (the individual who the template is based on), the participants, noise (a participant was selected at random to be observed in an environment which contained noise such as people walking in the background) and the average (a numerical representative of the overall respirations recorded, breaths detected, and the breaths missed).

Test Label	*Breath Recorded*	*Breath Detected*	*Breath missed*
Control	7	7	0
Participant 1	11	8	3
Participant 2	14	10	4
Participant 3	8	5	3
Noise	14	9	5
Average	**10**	**7**	**3**

Table 1: Run One – Shallow breath

Test Label	*Breath Recorded*	*Breath detected*	*Breath missed*
Control	14	12	2
Participant 1	10	9	1

Participant 2	16	13	3
Participant 3	11	9	2
Noise	12	9	3
Average	**12**	**10**	**2**

Table 2: Run Two – Normal breath

Test Label	*Breath Recorded*	*Breath detected*	*Breath missed*
Control	10	9	1
Participant 1	14	13	1
Participant 2	14	9	5
Participant 3	16	13	3
Noise	14	7	7
Average	**13**	**10**	**3**

Table 3: Run Three – Heavy breath

From the results presented in Table 1, 2 and 3, it can be seen that the proposed approach detects the majority of breaths using a single template on a range of individuals. It should be noted that shallow breathing was the most challenging action to detect because of fine and subtle movements, as shown in Table 1, and displayed the worst performance across the three tests when the detection results for each participant is individually analysed. Similarly heavy breathing (Table 3) suffered a similar average breath missed; shallow breaths possessed a lower number of recorded breaths, but the same number of breaths missed. During the distance phase, the best performing participant across the tests other than control (participant 1) was asked to repeat the same behaviour (shallow, normal and heavy breaths) as in the detection phase but with the added variable of distance of the participant respective to the neuromorphic vision hardware to determine how the approach handles different scales. The distances were 0.5, 1 and 1.5 metres for control. The results of the distance phase are provided in Tables 4, 5, 6, 7, 8 and 9. Tables 5, 7 and 9 show the results of the implemented approach observing the participant with noise from the environment as in the detection phase above.

Participant 1 Shallow Breathing Distance Recording			
Distance	Breath Recorded	Breath detected	Breath missed
0.5m	11	11	0
1m	9	7	2
1.5m	14	10	4

Table 4: Shallow Distance Run One

Participant 1 Shallow Breathing Distance Recording including Noise			
Distance	Breath Recorded	Breath detected	Breath missed
0.5m	10	7	3
1m	12	5	7
1.5m	10	4	6

Table 5: Shallow Distance Run Two

Participant 1 Normal Breathing Distance Recording			
Distance	Breath Recorded	Breath detected	Breath missed
0.5m	13	10	3
1m	13	7	6

1.5m	11	5	6

Table 6: Normal Distance Run One

Participant 1 Normal Breathing Distance Recording including Noise			
Distance	Breath Recorded	Breath detected	Breath missed
0.5m	14	10	4
1m	9	6	3
1.5m	12	5	7

Table 7: Normal Distance Run Two

Participant 1 Heavy Breathing Distance Recording			
Distance	Breath Recorded	Breath detected	Breath missed
0.5m	15	12	3
1m	18	15	3
1.5m	11	10	1

Table 8: Heavy Distance Run One

Participant 1 Heavy Breathing Distance Recording including Noise			
Distance	Breath Recorded	Breath detected	Breath missed
0.5m	10	6	4
1m	13	5	8
1.5m	15	4	11

Table 9: Heavy Distance Run Two

From the results presented in Tables 4, 5, 6, 7, 8 and 9, it is important to note that the proposed approach proceeds to become less effective as the distance from the vision sensor increased. This highlights a fault with the assumption that the same activity will always produce the same spike train. Although the same activity will produce similar spike trains (hence why the approach is still capable of recognising the action), when viewed at an increased distance from the camera, less events are produced because of the effects of apparent scale within the scene. This is an issue which is paramount within human vision systems where it increasingly difficult to identify fine grained details at a distance. The proposed approach, however, has the potential to be implemented inside a multi-scale framework which should allow for more versatility in the approach to the challenges of scale.

5 Conclusion

By taking advantage of the unique data type presented by neuromorphic vision sensors, such as the DVS128, the capability to detect an action using a template-based approach has been demonstrated in this paper. An approach to identify three different breathing actions (shallow, normal and heavy) using a sliding window templating architecture was developed and presented. The proposed approach was evaluated in terms of both distance and detection accuracy and we found that the approach is accurate for detection but can suffer degradation at scale. It can be hypothesised that the issue of degradation at scale can be overcome by implementing the approach inside a multi-scale framework, this will be explored as future work.

References

[Eberlein, 2015] Eberlein, P. (2015). Understanding Video Latency What is video latency and why do we care about it?. https://www.vision-systems.com/content/dam/VSD/solutionsinvision/Resources/Sensoray_video-

latency_article_FINAL.pdf

[Lichtsteiner, et al., 2008] Lichtsteiner, P., Posch, C., & Delbruck, T. (2008). A 128 x 128 120 dB 15 μs Latency Asynchronous Temporal Contrast Vision Sensor. *IEEE journal of solid-state circuits*, *43*(2), 566-576.

[Moeys, et al., 2016] Moeys, D. P., Corradi, F., Kerr, E., Vance, P., Das, G., Neil, D., & Delbrück, T. (2016). Steering a Predator Robot using a Mixed Frame. *Event-Driven Convolutional Neural Network Steering a Predator Robot using a Mixed Frame/Event-Driven Convolutional Neural Network.(July)*.

[Rayner, et al., 2009] Rayner, K., Smith, T. J., Malcolm, G. L., & Henderson, J. M. (2009). Eye movements and visual encoding during scene perception. *Psychological science*, *20*(1), 6-10.

[Shechtman, et al., 2005] Shechtman, E., & Irani, M. (2005, June). Space-time behavior based correlation. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* (Vol. 1, pp. 405-412). IEEE.

[Thorpe, et al., 2001] Thorpe, S., Delorme, A., & Van Rullen, R. (2001). Spike-based strategies for rapid processing. *Neural networks*, *14*(6-7), 715-725.

[Wiley, 1985] Wiley, D. G. (1985). Measuring reaction time. The Physics Teacher. 23(5). 314-314.